

Role of subgraphs in epidemics over finite-size networks under the scaled SIS process

June Zhang^{*} and José M.F. Moura[†]

Carnegie Mellon University

Electrical and Computer Engineering Dept.

Pittsburgh, PA, USA

(Dated: October 10, 2014)

Abstract

In previous work, we developed the scaled SIS process, which models the dynamics of SIS epidemics over networks. With the scaled SIS process, we can consider networks that are finite-sized and of arbitrary topology (i.e., we are not restricted to specific classes of networks). We derived for the scaled SIS process a closed-form expression for the time-asymptotic probability distribution of the states of all the agents in the network. This closed-form solution of the equilibrium distribution explicitly exhibits the underlying network topology through its adjacency matrix. This paper determines which network configuration is the most probable. We prove that, for a range of epidemics parameters, this combinatorial problem leads to a submodular optimization problem, which is *exactly* solvable in polynomial time. We relate the most-probable configuration to the network structure, in particular, to the existence of high density subgraphs. Depending on the epidemics parameters, subset of agents may be more likely to be infected than others; these more-vulnerable agents form subgraphs that are denser than the overall network. We illustrate our results with a 193 node social network and the 4941 node Western US power grid under different epidemics parameters.

PACS numbers: Valid PACS appear here

^{*} junez@andrew.cmu.edu

[†] Was a visiting professor with New York University and the Center for Urban Science and Policy (CUSP) in 2013-2014

I. INTRODUCTION

A network is a graph; it is a collection of nodes connected by edges. Networks have been used in science and engineering to represent systems of multiple interconnected, interdependent components. As a result, the network structure has a large impact on the behavior of the system. Quantifying how network structure impacts network function, that is, the behavior of dynamical processes on networks, is a difficult problem since the system components do not behave independently.

In this paper, we focus on analyzing the behavior of network diffusion processes such as epidemics. Analytical results for epidemics on networks have been obtained under particular conditions: full mixing models (i.e., the underlying network is a complete graph); infinite-sized networks models using mean-field approximation; or for scaled-free networks [1–4]. These approaches approximate the underlying network topology with mathematically simpler structures, because accounting for the exact graph topology is a combinatorial problem that is difficult to analyze and computationally expensive to compute. We showed in previous work [5, 6] that, for a specific network diffusion process, which we called the *scaled SIS* (Susceptible-Infected-Susceptible) process, it is possible to characterize its time-asymptotic behavior on any arbitrary, finite-sized network with N agents.

The scaled SIS process is Markov. It accounts for 1) exogenous (i.e., spontaneous) infection at rate λ ; 2) endogenous (i.e., neighbor-to-neighbor) infection at rate γ ; and 3) healing at rate μ . The time-asymptotic behavior of the process is described by its equilibrium distribution, which is a PMF (probability mass function) over all 2^N possible network configurations. Our approach preserves the full microscopic states of all the agents in contrast to previous approaches that only provide results for aggregate or macroscopic states (e.g., fraction of infected agents) [7]. However, retaining the exact network configuration means that the computational complexity of solving for the equilibrium distribution, an eigenvector problem, scales exponentially with the size of the network, N .

We have shown that, under specific assumptions on the form of the endogenous infection, the scaled SIS process is a *reversible* Markov process for which we can find its equilibrium distribution in closed form, avoiding solving a large eigenvalue/eigenvector problem. Further, the equilibrium distribution that we derived exhibits explicitly the underlying network structure through the network adjacency matrix. The equilibrium distribution is parame-

terized by two parameters: $\left(\frac{\lambda}{\mu}, \gamma\right)$, where as usual, parameter $\frac{\lambda}{\mu}$ controls the exogenous, or the topology-independent behavior of the scaled SIS process, whereas parameter γ controls the endogenous or the topology-dependent behavior of the process.

We used the equilibrium distribution to address the question of which of the 2^N possible configurations in a network is the most likely to occur in the long run. We refer to this as the most-probable configuration, which is found by maximizing the equilibrium distribution. This optimization (called the Most-Probable Configuration Problem) is difficult because: 1) it is combinatorial; 2) it depends on the healing/infection parameters of the scaled SIS process; and 3) it depends on the underlying network topology. Previously in [5], we partitioned the space of $\left(\frac{\lambda}{\mu}, \gamma\right)$ values into four regimes and were able to find the most-probable configuration in Regime II) **Endogenous Infection Dominant**, for which $0 < \frac{\lambda}{\mu} \leq 1, \gamma > 1$, for only specific types of networks: k -regular, complete multipartite, and complete multipartite with k -regular islands. We showed for these specific networks that the most-probable configuration solution space exhibits phase transition behavior depending on the network structure and epidemics parameters.

This paper considers the Most-Probable Configuration Problem for arbitrary networks. We are able to prove that this leads to the optimization of a submodular function for which we have a polynomial time solution. Further, we show which clusters of agents in the network are more vulnerable to epidemics than others. These are relevant questions in applications. For example, these are the clusters to focus on in marketing campaigns or when combating epidemics.

We review the scaled SIS process in Section II and set up the Most-Probable Configuration Problem in Section III. In Section IV, we show that, in Regime II), the Most-Probable Configuration Problem can be transformed into an equivalent submodular problem, and that it is possible to solve for its *exact* solution in *polynomial time*. We apply this to solve the most-probable configuration for two example networks: the 193 node acquaintance network of drug users in Hartford, CT [8], and the 4941 node network of the Western US power grid [9]. Section V shows how the solution space of the Most-Probable Configuration Problem in Regime II) relates to the density of subgraphs in the network. Section VI concludes the paper.

II. SCALED SIS PROCESS

Consider a population of N agents whose interconnections are represented by a static, simple, unweighted, undirected, connected graph, $G(V, E)$, where $V(G)$ is the set of vertices and $E(G)$ is the set of edges. For background on graphs see [10]. The topology of G is captured by the symmetric $N \times N$ adjacency matrix, A . The state of the i^{th} agent is denoted by x_i . Agents can be in one of two states: susceptible ($x_i = 0$) or infected ($x_i = 1$); susceptible agents are vulnerable to infections since there is no immunization in the system.

Let

$$\mathbf{x} = [x_1, x_2, \dots, x_N]^T.$$

We will refer to x_i as the *agent state* and \mathbf{x} as either the *network state* or the *network configuration*. The configuration state space is $\mathcal{X} = \{\mathbf{x}\}$, with cardinality $|\mathcal{X}| = 2^N$.

The scaled SIS process models the evolution of the network state, \mathbf{x} , over time according to the stochastic microscopic interaction rules from the SIS (susceptible-infected-susceptible) epidemics. The SIS framework assumes that infected agents can heal and become reinfected so it does not account for immunization [1]. Let $X(t) = \mathbf{x}$ be the state of the network at time t , $t \geq 0$. Under appropriate assumptions, $X(t)$ is a continuous-time Markov process [7, 11, 12]. The scaled SIS process accounts for 1) exogenous infection (i.e., susceptibles spontaneously develop infection); 2) endogenous infection (i.e., susceptibles become infected due to infection from infective neighbors); and 3) healing events. These processes are independent. At time t , only one agent is affected. By including both exogenous infection and healing, the scaled SIS process does *not* have an absorbing state at equilibrium.

The scaled SIS process is Markov; each network state is a state of the Markov process. We define two operators on the network state, $\mathbf{x} = [x_1, x_2, \dots, x_i, \dots, x_j, \dots, x_N]^T$. We use the following notation:

$$\begin{aligned} H_i \mathbf{x} &= [x_1, x_2, \dots, x_i = 1, \dots, x_N]^T \\ H_{j\bullet} \mathbf{x} &= [x_1, x_2, \dots, x_j = 0, \dots, x_N]^T. \end{aligned}$$

The operator H_i defines the operation that agent i becomes infected. If agent i is already infected, the operator does nothing. The operator $H_{j\bullet}$ defines the operation that agent j is healed. If agent j is already uninfected, the operator does nothing.

The time the process spends in a particular state is random and exponentially distributed,

with the following transition rates corresponding to infection and healing events, respectively:

1. $X(t)$ jumps to the network state where the i th agent, which was healthy, becomes infected with transition rate

$$q(\mathbf{x}, H_i \mathbf{x}) = \lambda \gamma^{d_i}, \quad \mathbf{x} \neq H_i \mathbf{x}, \quad (1)$$

where $d_i = \sum_{j=1}^N \mathbb{1}(x_j = 1) A_{ij}$, is the number of infected neighbors of node i . The symbol $\mathbb{1}(\cdot)$ is the indicator function, and $A = [A_{ij}]$ is the adjacency matrix of the arbitrary network G that captures the interactions among the agents. There are two components to the infection rate. If the i th agent has no infected neighbors, $d_i = 0$, and the transition rate reduces to $\lambda > 0$. We interpret λ as the exogenous infection rate, the rate a susceptible agent spontaneously becomes infected; it is the same for all the agents in the network. If the i th agent has d_i infected neighbors, the infective rate is $\lambda \gamma^{d_i}$; it is the product of λ and the endogenous infection rate, $\gamma > 0$, *scaled* by d_i , the number of infected neighbors of agent i . Because of this factor, the infective rate depends on the network topology.

2. $X(t)$ jumps to the network state where the j th agent, which was infected, heals with transition rate:

$$q(\mathbf{x}, H_{j\bullet} \mathbf{x}) = \mu, \quad \mathbf{x} \neq H_{j\bullet} \mathbf{x}. \quad (2)$$

The healing rate, $\mu > 0$, is the same for all the agents in the system.

A. Equilibrium Distribution

The evolution of the scaled SIS process is captured by the rate (infinitesimal) matrix \mathbf{Q} of the Markov process $X(t)$. The assumption that the underlying network G is connected assures that the Markov process is irreducible. Therefore, the equilibrium distribution, $\pi(\mathbf{x})$, exists and is given by the left eigenvector corresponding to the 0 eigenvalue of \mathbf{Q} , the rate matrix [13]. The problem in determining the equilibrium distribution $\pi(\mathbf{x})$ is that its computation is prohibitively expensive for meaningful sized networks since \mathbf{Q} is a $2^N \times 2^N$ matrix. This has limited the analysis of epidemics and spreading processes on networks to either: 1) full mixing models (e.g., where every agent comes in contact with every other

agents —the network is a complete graph); 2) to small scale simulations, where N is small so that $O((2^N)^3)$ operations are feasible; or 3) to mean field type approximations of special network configurations.

We proved in [6], see also [5], that the scaled SIS process is a *reversible* Markov process by showing that its equilibrium distribution satisfies not only the global balance equation but also the detailed balance equation [14]. For reversible Markov processes, the equilibrium distribution is unique. We derived the equilibrium distribution of the scaled SIS process to be:

$$\pi(\mathbf{x}) = \frac{1}{Z} \left(\frac{\lambda}{\mu} \right)^{1^T \mathbf{x}} \gamma^{\frac{\mathbf{x}^T A \mathbf{x}}{2}}, \quad \mathbf{x} \in \mathcal{X} \quad (3)$$

where Z is the partition function,

$$Z = \sum_{\mathbf{x} \in \mathcal{X}} \left(\frac{\lambda}{\mu} \right)^{1^T \mathbf{x}} \gamma^{\frac{\mathbf{x}^T A \mathbf{x}}{2}}. \quad (4)$$

Previous epidemics model call the ratio $\frac{\lambda}{\mu}$, the *effective infection rate* [15]. The equilibrium distribution, $\pi(\mathbf{x})$, factors as the product of three terms: 1) the normalization by the partition function; 2) the term $\left(\frac{\lambda}{\mu} \right)^{1^T \mathbf{x}}$ that is topology independent since the exogenous infection rate λ and the healing rate μ are identical for all the agents in the network, and the total number of infected agents, $1^T \mathbf{x}$, does not depend on the topology; and 3) the $\gamma^{\frac{\mathbf{x}^T A \mathbf{x}}{2}}$ that explicitly accounts for the exact network through its adjacency matrix A . It is topology dependent since the endogenous infection rate γ is scaled by the number of infected neighbors; the number of edges where both end nodes are infected (we call them *infected edges*), $\frac{\mathbf{x}^T A \mathbf{x}}{2}$, explicitly depends on the adjacency matrix of the underlying network.

B. Parameter Regimes

The scaled SIS Process can model different types of network diffusion processes depending on the values of the rate parameters; in particular, if the effective exogenous infection rate, $\frac{\lambda}{\mu}$, and the endogenous infection rate, γ , are between 0 and 1, or if they are greater than 1. In [5], we identified 4 regimes.

When both parameters are either between 0 and 1 or greater than 1, then the most-probable configuration is either the $\mathbf{x}^0 = [0, 0 \dots, 0]^T$ configuration or the $\mathbf{x}^N = [1, 1 \dots, 1]^T$

configuration. Reference [5] also investigated Regime III) where $\frac{\lambda}{\mu} > 1, 0 < \gamma \leq 1$. This regime models the counter-intuitive behavior where an increasing number of infected agents delays additional infection in the network. In this paper, we focus our analysis on Regime II) **Endogenous Infection Dominant:** $0 < \frac{\lambda}{\mu} \leq 1, \gamma > 1$. Regime II best models epidemics and similar types of spreading processes.

The effective exogenous infection rate, $\frac{\lambda}{\mu}$, indicates the preference of individual agents. With $0 < \frac{\lambda}{\mu} \leq 1$, the healing rate is larger than the exogenous infection rate; agents prefer the healthy state to the infected state. With $\gamma > 1$, however, additional infected neighbors increase the rate at which the healthy agent becomes infected; thereby the network helps to spread the infection. As a result, the network topology is crucial to determine the behavior of the scaled SIS process at equilibrium.

In the next section, we introduce the Most-Probable Configuration Problem, which solves for the network configurations with maximum equilibrium probability. Because there is *competition* between the topology independent term and the topology dependent term, the most-probable configuration exhibits complex phase transition behavior depending on the effective exogenous infection rate $\frac{\lambda}{\mu}$, the endogenous infection rate γ , and the underlying network topology.

III. MOST-PROBABLE CONFIGURATION PROBLEM

In the previous section, we showed that, for the scaled SIS process, we are able to derive its equilibrium distribution, $\pi(\mathbf{x})$, analytically, see equation (3). The equilibrium distribution describes the long-run behavior of the network epidemics. While the partition function (4) renders the exact calculation of the equilibrium distribution infeasible for meaningful size networks, knowing the equilibrium distribution expression allows us to quickly compare between network configurations, addressing, for example questions like which of the two is more probable. Of all the possible 2^N network configurations, one is of particular interest, namely, the configuration of infected and healthy agents that has a higher chance of occurring in the long run. This is the configuration \mathbf{x}^* that maximizes $\pi(\mathbf{x})$. Formally, \mathbf{x}^* maximizes the equilibrium probability:

$$\mathbf{x}^* = \arg \max_{\mathbf{x} \in \mathcal{X}} \pi(\mathbf{x}) = \arg \max_{\mathbf{x} \in \mathcal{X}} \left(\frac{\lambda}{\mu} \right)^{1^T \mathbf{x}} \gamma^{\frac{\mathbf{x}^T A \mathbf{x}}{2}}. \quad (5)$$

We call this the Most-Probable Configuration Problem and \mathbf{x}^* the *most-probable configuration*. The Most-Probable Configuration Problem is a combinatorial optimization problem as agents can only be in one of two states; its solution is dependent on the effective exogenous infection rate, $\frac{\lambda}{\mu}$, the endogenous infection rate γ , and the underlying network topology, captured by the adjacency matrix, A .

Previously in [5], we provided analytical results for the Most-Probable Configuration Problem in Regime II) **Endogenous Infection Dominant**: $0 < \frac{\lambda}{\mu} \leq 1, \gamma > 1$ for particular networks, namely, structured network topologies such as k -regular, complete multipartite, complete multipartite with k -regular islands. We observed a phase transition behavior. Below a threshold condition that depends on the parameters $(\frac{\lambda}{\mu}, \gamma)$ and on the network topology, the most-probable configuration is $\mathbf{x}^0 = [0, 0, \dots, 0]$, the configuration where all agents are susceptibles. Above the threshold condition, the most-probable configuration is $\mathbf{x}^N = [1, 1, \dots, 1]$, the configuration where all agents are infected.

This paper extends the analysis of the Most-Probable Configuration Problem in Regime II) to *arbitrary* network topologies. We will show that, for arbitrary networks, the most-probable configuration may be configurations other than \mathbf{x}^0 and \mathbf{x}^N . We call these solutions to the Most-Probable Configuration Problem *non-degenerate configurations*. These solutions are useful for identifying agents and communities that are more vulnerable to the epidemics. We will relate these communities to the structure of the networks in detail later. Figure 1 and Figure 2 show the most-probable configurations obtained by the method of Section IV for two example networks: a 193-node acquaintance network [8] and the 4941-node power grid [9]. These are non-degenerate configurations where only a subset of agents are infected.

In Section IV, we prove that we can solve *exactly* for the most-probable configuration in Regime II) in polynomial time using submodular optimization. Then, in Section V, we discuss the relationship between the most-probable configuration and the network topology, in particular, the relation between non-degenerate configurations and network topology.

IV. SUBMODULARITY AND THE MOST-PROBABLE CONFIGURATION

In this section, we solve the Most-Probable Configuration Problem in Regime II in polynomial time by showing that the problem can be transformed into a submodular function. First, we review the definition of submodular functions.

A. Submodular Function

The Most-Probable Configuration Problem is the maximization of a pseudo-Boolean function. Pseudo-Boolean functions are functions that map N binary variables to a real number [16]. Minimization of general pseudo-Boolean functions is NP-hard [17]. Grötschel, Lovász, and Schrijver, [18], proved that the minimization of a pseudo-Boolean function that is submodular can be done in polynomial time. If the function is supermodular, its maximization is in polynomial time.

A pseudo-Boolean function, $f : \{0, 1\}^N \rightarrow \mathcal{R}$, is also a set function $g : \mathcal{P}(V) \rightarrow \mathcal{R}$ where $\mathcal{P}(V)$ is the power set of $V = \{1, 2, \dots, N\}$. There are many equivalent definitions of submodularity [19]. The one we use in this paper is the following:

Definition IV.1 ([16]). *A set function, $g : \mathcal{P}(V) \rightarrow \mathcal{R}$, is submodular if and only if for any $\alpha_1 \subseteq V, \alpha_2 \subseteq \alpha_1, i \in V \setminus \alpha_1$:*

$$g(\alpha_1 \cup \{i\}) - g(\alpha_1) \leq g(\alpha_2 \cup \{i\}) - g(\alpha_2).$$

For a submodular function, the incremental gain of adding an element to the set α_1 is less than or equal to the gain of adding the element to a smaller subset of α_1 . A supermodular function has the inequality in the opposite direction.

B. Most-Probable Configuration: A Submodular Problem

The Most-Probable Configuration Problem (5) seeks the maximum of a pseudo-Boolean function that maps a 0-1 vector, the network configuration \mathbf{x} , to a scalar. The network configuration $\mathbf{x} \in \{0, 1\}^N$ is the characteristic vector or characteristic function of the set of infected agents: $\alpha_{\mathbf{x}} = \{i \mid i \in V, x_i = 1\}$. Let $h(\alpha_{\mathbf{x}})$ be the set of infected edges (i.e., edges where both end nodes are infected) in configuration \mathbf{x} : $h(\alpha_{\mathbf{x}}) = \{\{i, j\} \mid i, j \in V, x_i = 1, x_j = 1\}$.

The number of infected agents in configuration \mathbf{x} is $|\alpha_{\mathbf{x}}| = \mathbf{1}^T \mathbf{x}$. The number of infected edges is $|h(\alpha_{\mathbf{x}})| = \frac{\mathbf{x}^T A \mathbf{x}}{2}$. The Most-Probable Configuration Problem is then to solve for the maximum argument of

$$g(\alpha_{\mathbf{x}}) = \left(\frac{\lambda}{\mu}\right)^{|\alpha_{\mathbf{x}}|} \gamma^{|h(\alpha_{\mathbf{x}})|}. \quad (6)$$

We will prove in Theorem IV.2 that $-\log(g(\alpha_{\mathbf{x}}))$ is a submodular function. Therefore, we can solve for its minimum argument in polynomial time. Lemma IV.1 sets up some basic conditions that makes proving Theorem IV.2 easier.

Lemma IV.1. *Consider two sets of infected agents, $\alpha_1, \alpha_2 \subseteq V$ and $i \in V \setminus \alpha_1$. The cardinalities of α_1 and α_2 are $|\alpha_1| = n_1$ and $|\alpha_2| = n_2$, respectively; then $|\alpha_1 \cup \{i\}| = n_1 + 1$, and $|\alpha_2 \cup \{i\}| = n_2 + 1$. The numbers of infected edges induced by α_1 and α_2 are $|h(\alpha_1)| = e_1$ and $|h(\alpha_2)| = e_2$, respectively. Let $|h(\alpha_1 \cup \{i\})| = e_1 + m_1$ and $|h(\alpha_2 \cup \{i\})| = e_2 + m_2$; therefore m_1 is the number of additional infected edges created with the inclusion of agent i in α_1 and m_2 is the number of additional infected edges created with the inclusion of agent i in α_2 . Let $\alpha_2 \subseteq \alpha_1$. Then:*

1. $n_1 \geq n_2$.

2. $e_1 \geq e_2$.

3. $m_1 \geq m_2$.

Proof. 1. When $\alpha_2 \subset \alpha_1$, α_2 must have strictly fewer number of infected agents than α_1 . When $\alpha_2 = \alpha_1$, then they contain the same number of infected agents. Hence, $n_1 \geq n_2$.

2. When $\alpha_2 \subset \alpha_1$, infected agents in α_2 can not induce more infected edges than the number of infected edges induced by the infected agents in α_1 . When $\alpha_2 = \alpha_1$, then the infected agents in α_1 and α_2 will induce the same number of infected edges. Hence, $e_1 \geq e_2$.

3. Every infected agent in α_2 is an infected agent in α_1 . Every new infected edge connecting the infected agent $j \in \alpha_2$ with i is also a new infected edge in $\alpha_1 \cup \{i\}$. However, some edge may also have $j \in \alpha_1$. Hence, $m_1 \geq m_2$.

□

Theorem IV.2. *Let $g(\alpha_{\mathbf{x}})$ be the set function given in (6). If $\lambda > 0, \mu > 0$ and $\gamma \geq 1$, then $-\log(g(\alpha_{\mathbf{x}}))$ is a submodular function, where*

$$-\log(g(\alpha_{\mathbf{x}})) = -|\alpha_{\mathbf{x}}| \log\left(\frac{\lambda}{\mu}\right) - |h(\alpha_{\mathbf{x}})| \log(\gamma).$$

Proof. To prove submodularity of $-\log(g(\alpha_{\mathbf{x}}))$, we need to show that

$$-\log(g(\alpha_1 \cup \{i\})) + \log(g(\alpha_1)) \leq -\log(g(\alpha_2 \cup \{i\})) + \log(g(\alpha_2)), \quad (7)$$

for any $\alpha_1 \subseteq V, \alpha_2 \subseteq \alpha_1, i \in V \setminus \alpha_1$.

The left-hand side (LHS) of (7) is

$$-(n_1 + 1) \log\left(\frac{\lambda}{\mu}\right) - (e_1 + m_1) \log(\gamma) + n_1 \log\left(\frac{\lambda}{\mu}\right) + e_1 \log(\gamma), \quad (8)$$

which reduces to

$$-\log\left(\frac{\lambda}{\mu}\right) - m_1 \log(\gamma). \quad (9)$$

The right-hand side (RHS) of (7) is

$$-(n_2 + 1) \log\left(\frac{\lambda}{\mu}\right) - (e_2 + m_2) \log(\gamma) + n_2 \log\left(\frac{\lambda}{\mu}\right) + e_2 \log(\gamma), \quad (10)$$

which reduces to

$$-\log\left(\frac{\lambda}{\mu}\right) - m_2 \log(\gamma). \quad (11)$$

Expression (7) reduces to

$$-\log\left(\frac{\lambda}{\mu}\right) - m_1 \log(\gamma) \leq -\log\left(\frac{\lambda}{\mu}\right) - m_2 \log(\gamma).$$

Since $\gamma \geq 1$, we know that $\log(\gamma) \geq 0$ and that $m_1 \geq m_2$ by Lemma IV.1. Therefore, the LHS of (7) is less than or equal to the RHS of (7) for any $\alpha_1 \subseteq V, \alpha_2 \subseteq \alpha_1, i \in V \setminus \alpha_1$. By definition, $-\log(g(\alpha_{\mathbf{x}}))$ is a submodular function. \square

Theorem IV.2 proves that $-\log(g(\alpha_{\mathbf{x}}))$ is submodular if $\lambda > 0, \mu > 0$, and $\gamma \geq 1$; this means that $\log(g(\alpha_{\mathbf{x}}))$ is supermodular under the same condition. Since the logarithm function is a monotonic function, the maximum argument of $\log(g(\alpha_{\mathbf{x}}))$ is also the maximum argument of $g(\alpha_{\mathbf{x}})$, which is the solution to the Most-Probable Configuration Problem. As Regime II) **Endogenous Infection Dominant:** $0 < \frac{\lambda}{\mu} \leq 1, \gamma > 1$ satisfies the condition that $\gamma \geq 1$, using submodular optimization, we can find the *exact* most-probable configuration of the scaled SIS process in Regime II) for *arbitrary* network topology in polynomial time.

C. Social Networks and the Power Grid

The most-probable configuration allows us to identify the set of agents that are vulnerable to network epidemics since it retains the state of all the agents. Agents who are infected in the most-probable configuration are more vulnerable to the epidemics than agents who remain healthy. Because the most-probable configuration is derived from a dynamical model of network diffusion processes, the set of vulnerable agents depends on the infection and healing rates, λ, γ, μ .

As we showed in [5], the most-probable configuration changes depending on these parameters. When the healing rate is high, $\mathbf{x}^* = \mathbf{x}^0$, meaning that the epidemics is not severe. When the infection rate is high, $\mathbf{x}^* = \mathbf{x}^N$, the epidemics is severe, and all the agents are vulnerable. When \mathbf{x}^* is a non-degenerate configuration (i.e, $\mathbf{x}^* \neq \mathbf{x}^0, \mathbf{x}^N$), this indicates that sets of agents in the network are more vulnerable than others to the epidemics. We illustrate this by solving for the most-probable configuration using [20] under different $\left(\frac{\lambda}{\mu}, \gamma\right)$ parameters for 2 realistic networks: a social network [8] and the Western United States power grid [9], obtained from [21]

The network shown in Fig. 1 is a 193 node, 273 edge social network of drug users in Hartford, CT. The network was determined through interviews. Reference [22] looked for influential agents in the network by considering it as a graph connectivity problem. However, they did not consider a dynamical model of influence. Assuming that we can model drug habits as an epidemics (i.e., there is a social contagion aspect to the behavior), we applied the scaled SIS process to this network and solved for the most-probable configuration under different parameters to find influential network structures.

We show the resultant most-probable configurations in Fig. 1a, Fig. 1b, Fig. 1c, Fig. 1d as we change $\left(\frac{\lambda}{\mu}, \gamma\right)$. We can see from these results that there is a small community of users who are infected when others are healthy. The size of this community increases or decreases depending on the parameters. If there is a social contagion component to drug usage, then these agents may be more vulnerable to the social contagion component of drug usage and therefore more likely to persist in their habit. In the next section, we will relate the most-probable configuration to the network substructure.

The network shown in Fig. 2 is the 4941 node, 6595 edge power grid network of the Western United States used by Watts and Strogatz. They showed through simulation of the

SIR (susceptible-infected-removed) epidemics model on the western power grid that small-world networks like the western power grid are more conducive to spreading infection/failures than lattice networks. This is useful for explaining why failures propagate so quickly in a blackout. However, they did not identify *which* components in the power grid are more vulnerable to the epidemics.

Figure 2a and Fig. 2b show the most-probable configuration for the western US power grid when for the scaled SIS process parameterized $\left(\frac{\lambda}{\mu} = 0.33, \gamma = 2\right)$ and $\left(\frac{\lambda}{\mu} = 0.33, \gamma = 2.6\right)$, respectively. We can see that for the same $\frac{\lambda}{\mu}$, as γ increases, thereby increasing the infectiousness of epidemics, the number of infected agents increases. This is intuitive since, for large γ , the epidemics is severe, and the most-probable configuration is driven toward \mathbf{x}^N , the configuration where all the agents are infected. Moreover, the most-probable configurations are both non-degenerate configurations. The agents who are infected at equilibrium are more vulnerable to the network epidemics than agents who are healthy. By using submodular optimization, we can identify these more vulnerable agents, by solving for the most-probable configuration out of 2^{4941} total possible configurations, exactly and in polynomial time.

An important question is to relate the most-probable configuration to network structure. We will show in the next section that the most-probable configuration is related to subgraph density by rewriting the equilibrium distribution (3) in terms of induced subgraphs instead of network configurations.

V. MOST-PROBABLE CONFIGURATION AND NETWORK STRUCTURE

In the previous section, we showed that we can exactly solve for the most-probable configuration with a polynomial time algorithm. The exact solution, however, does not give insight on how the most-probable configuration changes depending on the parameters $\left(\frac{\lambda}{\mu}, \gamma\right)$ and on the network topology. In this section, we draw the connection between the most-probable configuration and subgraphs in the network. As per our intuition for epidemics, densely connected network structures are more vulnerable to network epidemics; the scaled SIS process quantifies this intuition. First, we will define the graph theoretic terms used in this section.

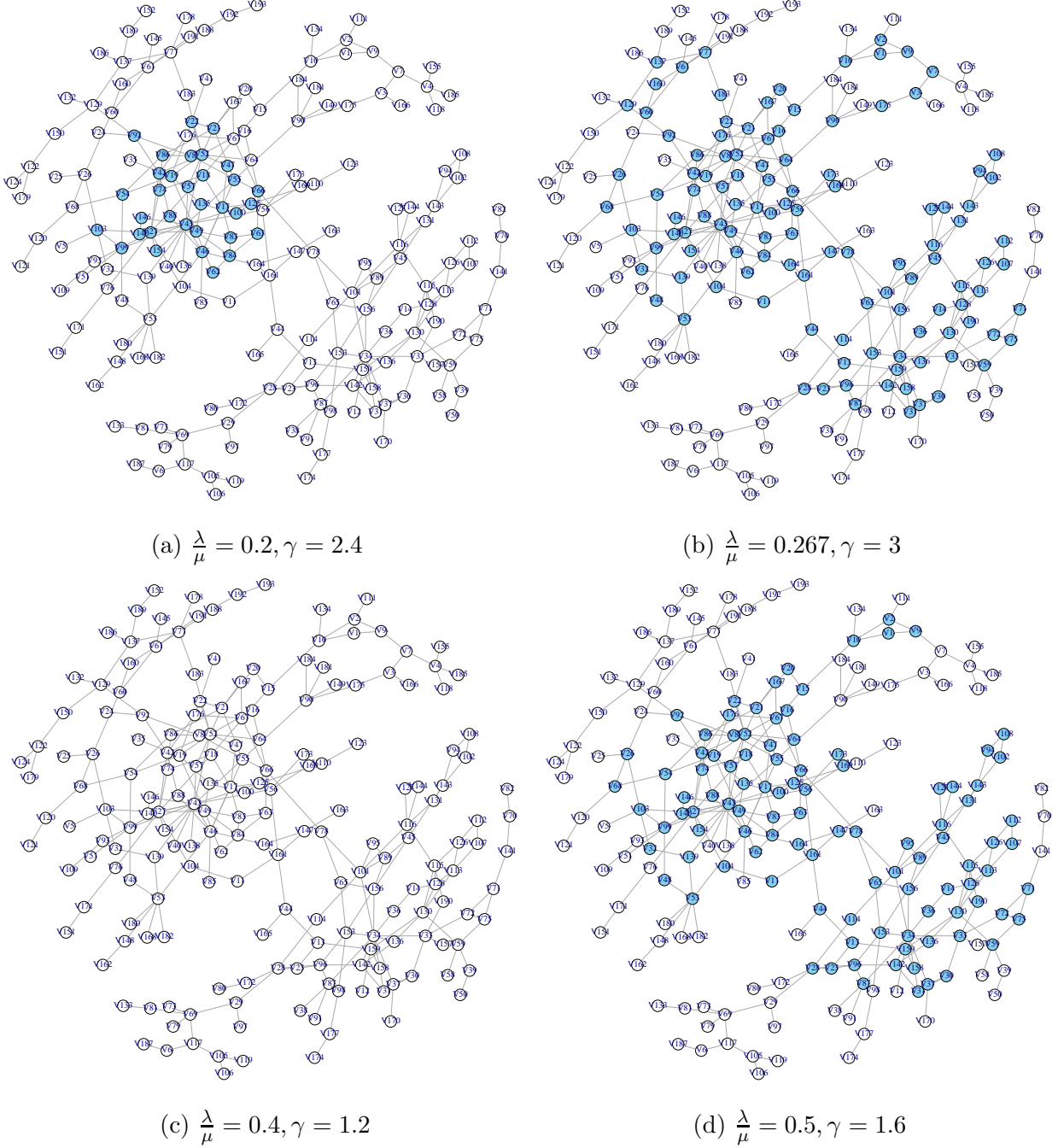
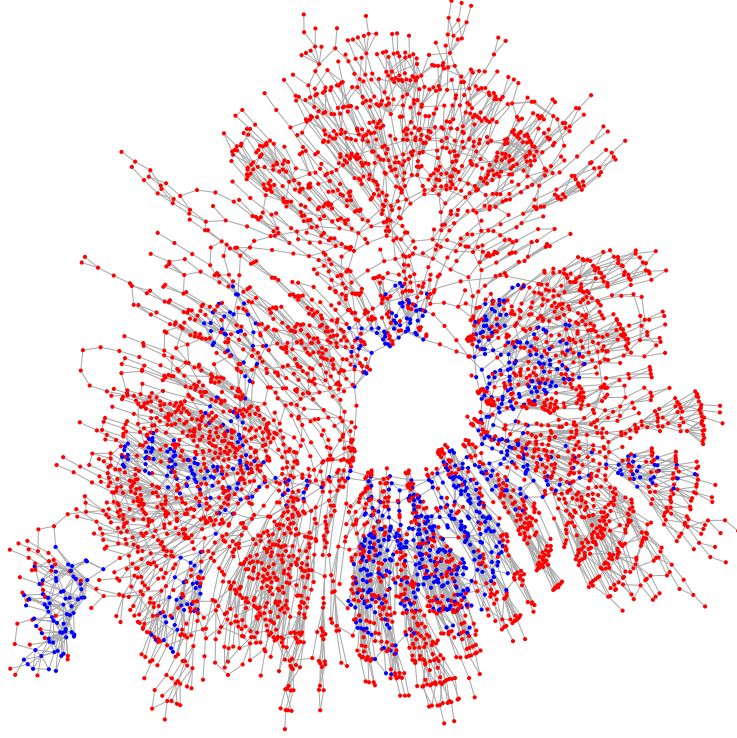


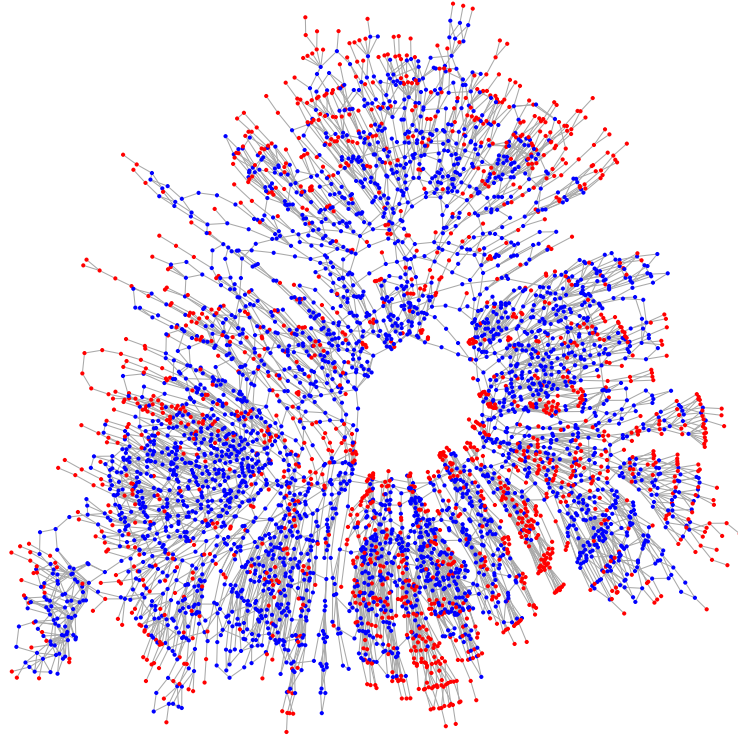
FIG. 1: (Color online) Most-Probable Configuration \mathbf{x}^* under Different $\left(\frac{\lambda}{\mu}, \gamma\right)$ Parameters (Blue/Grey = Infected, White = Healthy).

A. Induced Subgraphs and Graph Density

Definition V.1 (From [10]). *The graph H is an induced subgraph of G if two vertices in H are connected if and only if they are connected in G and the vertex set and edge set of H*



(a) $\frac{\lambda}{\mu} = 0.33, \gamma = 2$



(b) $\frac{\lambda}{\mu} = 0.33, \gamma = 2.6$

FIG. 2: (Color online) Most-Probable Configuration \mathbf{x}^* under Different $\left(\frac{\lambda}{\mu}, \gamma\right)$ Parameters (Blue/Black = Infected, Red = Healthy).

are subsets of the vertex set and edge set of G .

$$V(H) \subseteq V(G), E(H) \subseteq E(G)$$

Definition V.2. The graph $H(\mathbf{x})$ is an induced subgraph of configuration $\mathbf{x} = [x_1, x_2, \dots, x_N]^T$ if the nodes/edges in the subgraph are the infected agents/edges in \mathbf{x} .

$$V(H(\mathbf{x})) = \{v_i \in V(G) \mid x_i = 1\} \quad (12)$$

$$E(H(\mathbf{x})) = \{(i, j) \in E(G) \mid x_i = 1, x_j = 1\} \quad (13)$$

By definition, $|V(H(\mathbf{x}))| = \mathbf{1}^T \mathbf{x}$ and $|E(H(\mathbf{x}))| = \frac{\mathbf{x}^T A \mathbf{x}}{2}$. Figure 3 and Fig. 4 show two network configurations and their corresponding induced subgraphs. We proved in [23] that configurations whose induced subgraphs are isomorphic are equally probable. Unless we need to refer explicitly to the underlying network configuration \mathbf{x} , for notational simplicity, we will write H to denote an induced subgraph instead of writing $H(\mathbf{x})$.

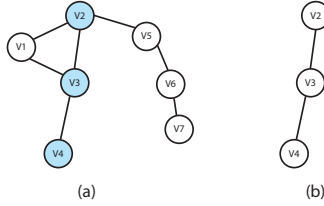


FIG. 3: (a) Configuration $\mathbf{x}_1 = [0, 1, 1, 1, 0, 0, 0]^T$, (b) Induced Subgraph $H(\mathbf{x}_1) = H_1$

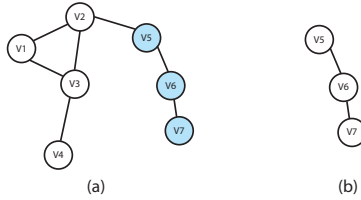


FIG. 4: (a) Configuration $\mathbf{x}_2 = [0, 0, 0, 0, 1, 1, 1]^T$, (b) Induced Subgraph $H(\mathbf{x}_2) = H_2$

Definition V.3. The set of all possible induced subgraphs of G is $\mathcal{H} = \{H(\mathbf{x})\}$, $\forall \mathbf{x} \in \mathcal{X}$.

The set \mathcal{H} includes the empty graph, which is induced by the configuration $\mathbf{x}^0 = [0, 0, \dots, 0]^T$, and G , which is the subgraph induced by the configuration $\mathbf{x}^N = [1, 1, \dots, 1]^T$.

Definition V.4 (From [24]). *The density of the graph G is*

$$d(G) = \frac{|E(G)|}{|V(G)|}.$$

There is an alternative definition for graph density that is the number of edges divided by the total number of possible edges [25]. Unfortunately, these two definitions of density are not equivalent.

We will refer to the density of the entire network, $d(G) = d(H(\mathbf{x}^N))$, as the *network density*, and the density of an induced subgraph of G as the *subgraph density*. The density of the empty graph, $d(H(\mathbf{x}^0))$, is 0 by definition. The subgraphs in \mathcal{H} can be partially ordered by their density. There may be many subgraphs with the same density. A special induced subgraph in \mathcal{H} is the densest subgraph.

Definition V.5. *Let \overline{H} be the densest subgraph in G . Then*

$$d(\overline{H}) \geq d(H), \quad \forall H \in \mathcal{H}.$$

Finding \overline{H} is known as the *Densest Subgraph Problem*. It is known that this problem can be solved in polynomial time exactly and in linear time in approximation for undirected graphs [24].

B. Equilibrium Distribution of the Scaled SIS Process

Since there is a one-to-one relationship between the network configuration \mathbf{x} and its induced subgraph $H(\mathbf{x})$, we can rewrite the equilibrium distribution (3) of the scaled SIS process in terms of the induced subgraph density and the size of the induced subgraph:

$$\pi(H) = \frac{1}{Z} \left(\left(\frac{\lambda}{\mu} \right) \gamma^{d(H)} \right)^{|V(H)|}, \quad H \in \mathcal{H}, \quad (14)$$

where $d(H)$ is the density of the subgraph and Z is the partition function.

The Most-Probable Configuration Problem (5) is then also an optimization problem over all the possible induced subgraphs in G :

$$H(\mathbf{x}^*) = \arg \max_{H \in \mathcal{H}} \left(\left(\frac{\lambda}{\mu} \right) \gamma^{d(H)} \right)^{|V(H)|}. \quad (15)$$

The subgraph induced by the most-probable configuration, $H(\mathbf{x}^*)$, is the *most-probable subgraph*, but this is *not* necessarily the same subgraph as the densest subgraph, \overline{H} .

Stating the equilibrium distribution in terms of the induced subgraph will allow us to derive several theorems regarding the most-probable configuration. For the theorems that follow, we make the following assumptions:

Assumption 1.: The scaled SIS process operates in Regime II) **Endogenous Infection Dominant**. This limits the effective infection and the endogenous infection to the range, $0 < \frac{\lambda}{\mu} \leq 1$ and $\gamma > 1$.

Assumption 2.: The underlying network G is a simple, undirected, unweighted, and connected graph.

C. Most-Probable Configuration and Subgraphs

Theorem V.6. *[Proof in Appendix A] The most-probable configuration $\mathbf{x}^* \neq \mathbf{x}^0$ if and only if there exists at least one induced subgraph $H \in \mathcal{H}$ with density $d(H)$ for which $\lambda\gamma^{d(H)} > \mu$.*

Theorem V.7. *[Proof in Appendix B]*

Case 1: The densest subgraph, \overline{H} , is the network G . Then, $\mathbf{x}^ \neq \mathbf{x}^N$ if and only if $\frac{\lambda}{\mu}\gamma^{d(G)} \leq 1$.*

Case 2: The densest subgraph, \overline{H} , is not the network G . Then, $\mathbf{x}^ \neq \mathbf{x}^N$ if and only if there exists at least one induced subgraph $H \in \mathcal{H} \setminus G$ with density $d(H) = \frac{E'}{N'}$ for which*

$$\frac{\log(\frac{\lambda}{\mu}\gamma^{d(G)})}{\log(\frac{\lambda}{\mu}\gamma^{d(H)})} < \frac{N'}{N}. \quad (16)$$

Corollary V.8. *[Proof in Appendix C] Let the density of the network be $d(G) = \frac{E}{N}$. Then, the most-probable configuration is a non-degenerate configuration, $\mathbf{x}^* \in \mathcal{X} \setminus \{\mathbf{x}^0, \mathbf{x}^N\}$, if and only if there exists at least one induced subgraph $H \in \mathcal{H}$ with density $d(H) = \frac{E'}{N'}$ for which $\lambda\gamma^{d(H)} > \mu$, and*

$$\frac{\log(\frac{\lambda}{\mu}\gamma^{d(G)})}{\log(\frac{\lambda}{\mu}\gamma^{d(H)})} < \frac{N'}{N}.$$

In Regime II) individual agents have a preference for being healthy, but the epidemics might spread to other agents through neighbor-to-neighbor contagion. Under the scaled SIS process, the subgraph density $d(H)$ scales the exogenous infection rate γ , thereby affecting the overall infection rate. Theorem V.6 states that, if the network contains *dense-enough*

subgraphs, then even when the effective exogenous infection rate, $\frac{\lambda}{\mu}$, is small (i.e., $0 < \frac{\lambda}{\mu} \ll 1$), the exogenous infection rate, γ , can leverage dense subgraphs to spread the infection throughout the network.

On the other hand, if the endogenous infection rate, γ , is large (i.e., $\gamma \gg 1$), then most certainly the epidemics will spread throughout the entire network. Theorem V.7 states when this does not happen. Furthermore, Theorem V.7 shows that it is important to consider if the densest subgraph in the network is the entire network or a smaller subgraph. Corollary V.8 proves that the existence of the non-degenerate configurations is related to the existence of subgraphs with density larger than the network density. The existence of these *denser-than* G subgraphs is crucial to the existence of non-degenerate configurations (i.e., different from \mathbf{x}^0 and \mathbf{x}^N) as solutions to the Most-Probable Configuration Problem; when the most-probable configuration is a non-degenerate configuration, agents belonging to denser subgraphs are more vulnerable to the epidemics.

In network science, dense clusters of agents have often been identified as either the network *core* or *community* [2, 26, 27]. Solving for the non-degenerate configuration is an alternative method for determining these network structures. Previous works in core/community detection are algorithmic and do not consider the dynamical process on the network. The scaled SIS process, however, is a model for dynamical processes on networks and, therefore, what is considered a *community* changes depending on the parameters of the dynamical process: the most-probable configuration changes depending on the exogenous rates $\frac{\lambda}{\mu}$ and on the endogenous rates γ .

To get an easy visual interpretation of Theorem V.6 and V.7, we illustrate them with two small 16-node examples; Network A shown in Fig. 5 and Network B in Fig. 6. For each network, we fix the effective exogenous infection rate, $\frac{\lambda}{\mu} = 0.5$. We then solve for the most-probable configuration for different γ , ranging from 1.2 to 3. As the endogenous infection rate, γ , changes, the most-probable configuration also changes. In Fig. 5a and Fig. 6a, neither network supports dense enough subgraphs for the epidemics to spread. But as γ increases, the infection starts to spread. In Network A, there is at least one subgraph denser than the network. The subgraph induced by $V1, V2, V3, V4, V5, V7, V8, V9, V10$ has a density of 1.33 whereas the density of the entire network is 1.19. In Fig. 5b, the most-probable configuration has these 9 agents infected while the other 7 agents remain healthy. The 9 agents in the dense subgraph are more vulnerable to the epidemics when $\frac{\lambda}{\mu} = 0.5$ and

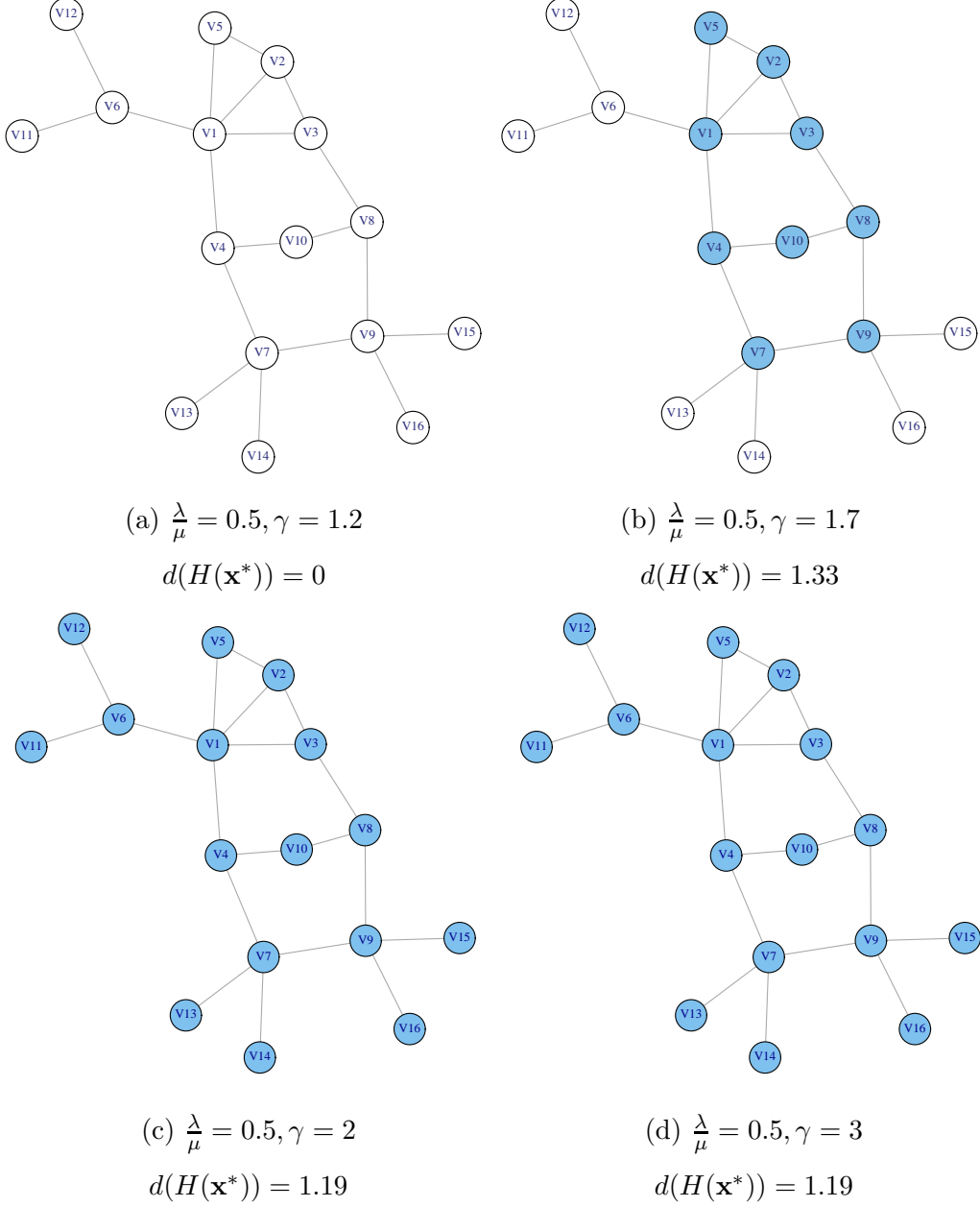


FIG. 5: (Color online) Most-Probable Configuration \mathbf{x}^* under Different $\left(\frac{\lambda}{\mu}, \gamma\right)$ Parameters (Blue/Grey = Infected, White = Healthy).

$\gamma = 1.7$.

In Network B, there are at least two subgraphs denser than the network and they are induced by the set of infected agents of the most-probable configuration as shown in Fig. 6b and Fig. 6c. We can see by solving for the most-probable configuration for different parameter values that, as the endogenous infection increases, the most-probable configuration goes

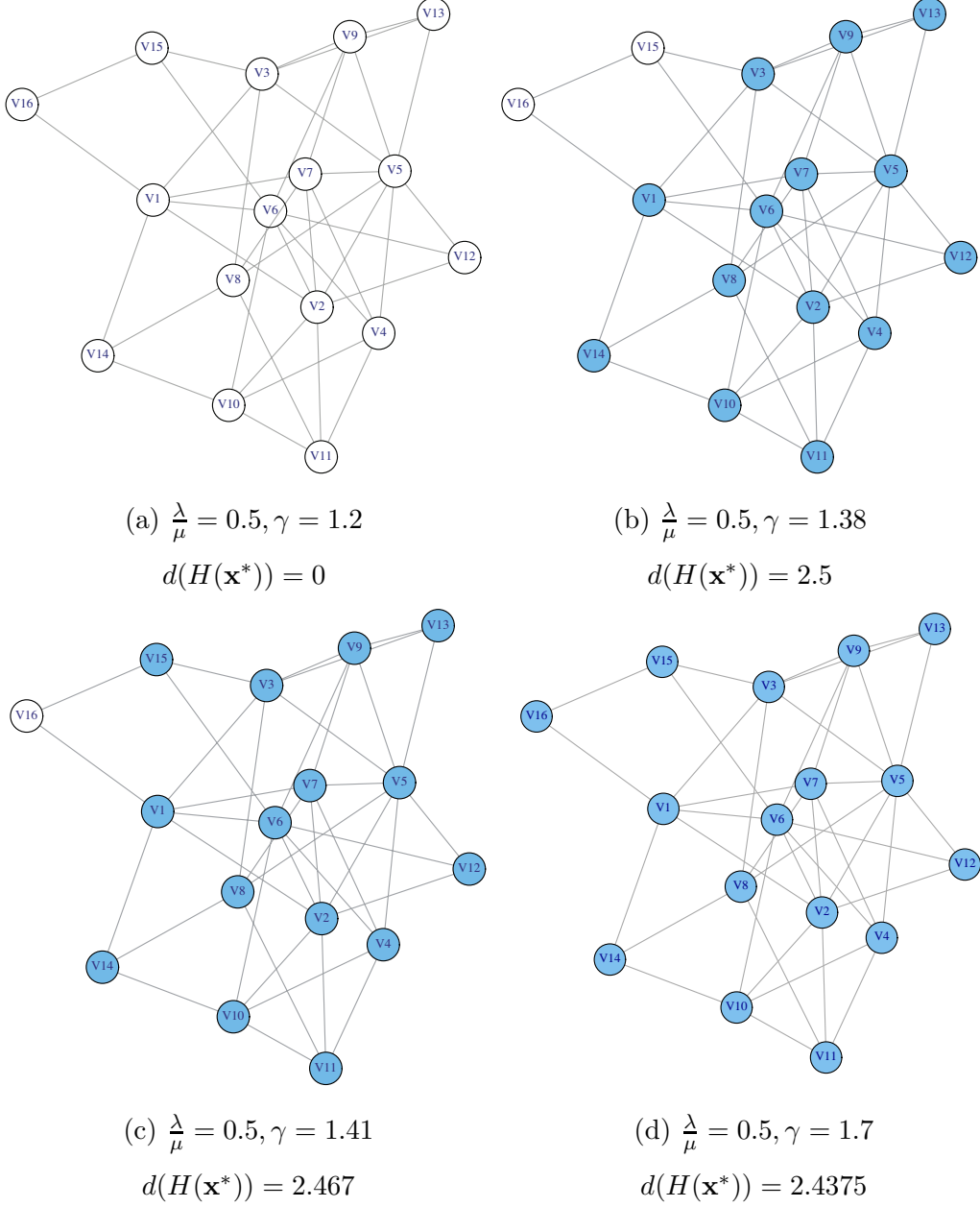


FIG. 6: (Color online) Most-Probable Configuration \mathbf{x}^* under Different $\left(\frac{\lambda}{\mu}, \gamma\right)$ Parameters (Blue/Grey = Infected, White = Healthy).

toward \mathbf{x}^N as all agents become vulnerable to the epidemics.

It is easier for the infection to spread in Network B than in Network A, since, at the same effective exogenous infection rate, $\mathbf{x}^* = \mathbf{x}^N$ for Network B when $\gamma = 1.7$ but $\mathbf{x}^* \neq \mathbf{x}^N$ for Network A with the same exogenous infection rate. This is because Network B is a denser graph ($d(G) = 2.4375$) than Graph A ($d(G) = 1.19$).

D. Most-Probable Configuration and the Densest Subgraph

We showed that the most-probable configuration is related to the density of induced subgraphs in the network. The densest subgraph, \overline{H} , is a special induced subgraph. In this section, we focus specifically on the relationship between the most-probable configuration and the densest subgraph.

Corollary V.9. *[Proof in Appendix D] The most-probable configuration $\mathbf{x}^* = \mathbf{x}^0$ if and only if $\lambda\gamma^{d(\overline{H})} \leq \mu$.*

Corollary V.9 follows the result of Theorem V.6. If the densest subgraph in the network is not *dense enough* to overcome individual preferences for being healthy, then the endogenous infection rate γ will not be able to drive the most-probable configuration away from \mathbf{x}^0 .

Lastly, because of the connection between the most-probable configuration of the scaled SIS process and the densest subgraph, we can prove a *general* statement regarding network structure using results from dynamical processes on networks.

Corollary V.10. *[Proof in Appendix E] If G is a k -regular, complete multipartite, or complete multipartite with k -regular islands network, then $\overline{H} = G$. That is, for these structured networks, the densest subgraph is the overall graph.*

VI. CONCLUSION

We introduced in previous works the scaled SIS process, which is a mathematically analyzable model for modeling diffusion processes on a static network [5]. The scaled SIS process is a reversible Markov process and has a closed-form equilibrium distribution that explicitly accounts for the underlying network topology via the adjacency matrix. It is parameterized by 2 parameters: $(\frac{\lambda}{\mu}, \gamma)$. The effective exogenous infection rate $\frac{\lambda}{\mu}$ controls the exogenous, or the topology-independent behavior of the scaled SIS process whereas the exogenous infection rate γ controls the endogenous or the topology-dependent behavior of the process.

Depending on if the parameter values are between 0 and 1 or great than 1, the scaled SIS process models qualitatively different network diffusion processes. In Regime II) **Endogenous Infection Dominant:** $0 < \frac{\lambda}{\mu} \leq 1, \gamma > 1$, the scaled SIS process best models

a network epidemics process; individuals prefer to be healthy, while neighbor-to-neighbor infection helps to spread the epidemics throughout the population.

This paper analyzes the Most-Probable Configuration Problem that solves for the network state with the maximum equilibrium probability, in Regime II) for arbitrary networks. First, we prove that the Most-Probable Configuration Problem in Regime II) is submodular. This means that we can compute the *exact* most-probable configuration in *polynomial time*. We use the most-probable configuration of the scaled SIS process to identify sets of vulnerable agents/components for a social network of drug users and the Western US power grid under different infection/healing rates.

We then showed that the most-probable configuration is dependent on certain classes of subgraphs in the networks. If there exist *dense-enough* subgraphs, conditioned on the right set of parameters, the most-probable configuration will shift away from \mathbf{x}^0 , the network state where all the agents are healthy. However, if there exist subgraphs that are *denser-than* the entire network, conditioned on the right range of infection and healing rates, the most-probable configuration may not reach \mathbf{x}^N , the network state with all agents infected. We call the solution of the Most-Probable Configuration Problem that is neither \mathbf{x}^0 nor \mathbf{x}^N , the non-degenerate configuration. Non-degenerate configurations identify subsets of agents that are more vulnerable to the network epidemics than others.

We also proved in this paper using results in [5] that structured networks such as k -regular, complete multipartite, complete multipartite with k -regular islands do not contain subgraphs that are denser than the overall network. Therefore, if we want to avoid subsets of agents being more vulnerable than others, we should use these types of structured networks.

Our analysis of the scaled SIS process in Regime II) informs us that network subgraph structures are important for understanding network diffusion processes. For future work, we are interested in statistically characterizing the subgraphs in network classes such as small-world networks and scaled-free networks.

ACKNOWLEDGMENTS

This work is partially supported by AFOSR grant FA95501010291, and by NSF grants CCF1011903 and CCF1018509. We wish to thank Prof. João P. Costeira and Prof. João M.F. Xavier of the Department of Electrical and Computer Engineering at Instituto Superior

Técnico, Lisbon, Portugal, for discussions regarding submodular optimization.

Appendix A: Proof for Theorem V.6

Theorem. *The most-probable configuration $\mathbf{x}^* \neq \mathbf{x}^0$ if and only if there exists at least one induced subgraph $H \in \mathcal{H}$ with density $d(H)$ for which $\lambda\gamma^{d(H)} > \mu$.*

Proof. Sufficiency: If there exists at least one subgraph $H \in \mathcal{H}$ with density $d(H)$ for which $\lambda\gamma^{d(H)} > \mu$, then $\mathbf{x}^* \neq \mathbf{x}^0$.

Using the equilibrium distribution (3), $\pi(\mathbf{x}^0) = \frac{1}{Z}$. Let the subgraph $H \in \mathcal{H}$ be the subgraph induced by configuration $\mathbf{x}' \in \mathcal{X} \setminus \mathbf{x}^0$. The number of infected agents in configuration \mathbf{x}' is $1^T \mathbf{x}' = |V(H)| > 0$. Using (14), its equilibrium probability is

$$\pi(\mathbf{x}') = \pi(H) = \frac{1}{Z} \left(\left(\frac{\lambda}{\mu} \right) \gamma^{d(H)} \right)^{|V(H)|}$$

If $\left(\frac{\lambda}{\mu} \right) \gamma^{d(H)} > 1$, we know that $\pi(\mathbf{x}') > \pi(\mathbf{x}^0)$. Therefore, \mathbf{x}^0 can not be the most-probable configuration.

Necessity: If $\mathbf{x}^* \neq \mathbf{x}^0$, then there exist at least one subgraph $H \in \mathcal{H}$ with density $d(H)$ for which $\lambda\gamma^{d(H)} > \mu$.

If $\mathbf{x}^* \neq \mathbf{x}^0$, this means that there is some configuration \mathbf{x}' for which $\pi(\mathbf{x}') > \pi(\mathbf{x}^0)$. We know that $\pi(\mathbf{x}^0) = \frac{1}{Z}$. Using the equilibrium distribution in (14) and the fact that $1^T \mathbf{x} = |V(H)| > 0, \forall \mathbf{x} \in \mathcal{X} \setminus \mathbf{x}^0$, we can conclude that there must exist some induced subgraph whose density satisfies this condition $\left(\frac{\lambda}{\mu} \right) \gamma^{d(H(\mathbf{x}'))} > 1$. \square

Appendix B: Proof for Theorem V.7

Theorem. *Case 1: The densest subgraph, \overline{H} , is the network G . Then, $\mathbf{x}^* \neq \mathbf{x}^N$ if and only if $\frac{\lambda}{\mu} \gamma^{d(G)} \leq 1$.*

Case 2: The densest subgraph, \overline{H} , is not the network G . Then, $\mathbf{x}^ \neq \mathbf{x}^N$ if and only if there exists at least one induced subgraph $H \in \mathcal{H} \setminus G$ with density $d(H) = \frac{E'}{N'}$ for which*

$$\frac{\log\left(\frac{\lambda}{\mu} \gamma^{d(G)}\right)}{\log\left(\frac{\lambda}{\mu} \gamma^{d(H)}\right)} < \frac{N'}{N}. \quad (\text{B1})$$

Proof. Sufficiency: Lets first prove sufficiency for both case 1 and case 2.

Case 1: $\overline{H} = G$. If $\lambda\gamma^{d(G)} \leq \mu$, then $\mathbf{x}^* \neq \mathbf{x}^N$.

Follows from Corollary V.9: If $\lambda\gamma^{d(\overline{H}(\mathbf{x}))} \leq \mu$, then $\mathbf{x}^* = \mathbf{x}^0$.

Case 2: $\overline{H} \neq G$. If there exists at least one induced subgraph $H \in \mathcal{H}$ with density $d(H) = \frac{E'}{N'}$ such that $\frac{\log(\frac{\lambda}{\mu}\gamma^{d(G)})}{\log(\frac{\lambda}{\mu}\gamma^{d(H)})} < \frac{N'}{N}$, then $\mathbf{x}^* \neq \mathbf{x}^N$.

The subgraph H is induced by the configuration $\mathbf{x}' \in \mathcal{X}$. The log equilibrium probability according to (14) for \mathbf{x}' and \mathbf{x}^N , respectively, are:

$$\log(\pi(\mathbf{x}')) = \log\left(\frac{1}{Z}\right) + N' \log\left(\frac{\lambda}{\mu}\gamma^{d(H)}\right)$$

and

$$\log(\pi(\mathbf{x}^N)) = \log\left(\frac{1}{Z}\right) + N \log\left(\frac{\lambda}{\mu}\gamma^{d(G)}\right).$$

Condition $\frac{\log(\frac{\lambda}{\mu}\gamma^{d(G)})}{\log(\frac{\lambda}{\mu}\gamma^{d(H)})} < \frac{N'}{N}$ implies that $N \log\left(\frac{\lambda}{\mu}\gamma^{d(G)}\right) < N' \log\left(\frac{\lambda}{\mu}\gamma^{d(H)}\right)$. Therefore, $\log(\pi(\mathbf{x}')) > \log(\pi(\mathbf{x}^N))$. Since the logarithm is a monotonic function, we can conclude that $\mathbf{x}^* \neq \mathbf{x}^N$.

Necessity: We now prove necessity for both case 1 and case 2.

Case 1: $\overline{H} = G$. If $\mathbf{x}^* \neq \mathbf{x}^N$, then $\lambda\gamma^{d(G)} \leq \mu$.

Follows from Corollary V.9: If $\mathbf{x}^* = \mathbf{x}^0$, then $\lambda\gamma^{d(\overline{H}(\mathbf{x}))} \leq \mu$.

Case 2: $\overline{H} \neq G$. If $\mathbf{x}^* \neq \mathbf{x}^N$, then there exists at least one induced subgraph $H \in \mathcal{H}$ such that $\frac{\log(\frac{\lambda}{\mu}\gamma^{d(G)})}{\log(\frac{\lambda}{\mu}\gamma^{d(H)})} < \frac{N'}{N}$.

Let $\mathbf{x}^* = \mathbf{x}'$, which induces a subgraph $H \in \mathcal{H}$ with density $d(H)$. Using (14),

$$\pi(\mathbf{x}') = \log\left(\frac{1}{Z}\right) + N' \log\left(\frac{\lambda}{\mu}\gamma^{d(H)}\right)$$

$$\pi(\mathbf{x}^N) = \log\left(\frac{1}{Z}\right) + N \log\left(\frac{\lambda}{\mu}\gamma^{d(G)}\right).$$

This means $\pi(\mathbf{x}') - \pi(\mathbf{x}^N) > 0$, which implies

$$N' \log\left(\frac{\lambda}{\mu}\gamma^{d(H)}\right) - N \log\left(\frac{\lambda}{\mu}\gamma^{d(G)}\right) > 0$$

This reduces to the condition that

$$\frac{\log(\frac{\lambda}{\mu}\gamma^{d(G)})}{\log(\frac{\lambda}{\mu}\gamma^{d(H)})} < \frac{N'}{N}.$$

□

Appendix C: Proof for Corollary V.8

Corollary. *Let the density of the network be $d(G) = \frac{E}{N}$. Then, the most-probable configuration is a non-degenerate configuration, $\mathbf{x}^* \in \mathcal{X} \setminus \{\mathbf{x}^0, \mathbf{x}^N\}$, if and only if there exists at least one induced subgraph $H \in \mathcal{H}$ with density $d(H) = \frac{E'}{N'}$ for which $\lambda\gamma^{d(H)} > \mu$, and*

$$\frac{\log(\frac{\lambda}{\mu}\gamma^{d(G)})}{\log(\frac{\lambda}{\mu}\gamma^{d(H)})} < \frac{N'}{N}.$$

Proof. Theorem V.6 gives the necessary and sufficient condition for the most-probable configuration $\mathbf{x}^* \neq \mathbf{x}^0$ to be existence of a subgraph H such that $\lambda\gamma^{d(H)} > \mu$. Theorem V.7 gives the necessary and sufficient condition that the most-probable configuration is not \mathbf{x}^N when

$$\frac{\log(\frac{\lambda}{\mu}\gamma^{d(G)})}{\log(\frac{\lambda}{\mu}\gamma^{d(H)})} < \frac{N'}{N}.$$

This proves the Corollary. □

Appendix D: Proof for Corollary V.9

Corollary. *The most-probable configuration $\mathbf{x}^* = \mathbf{x}^0$ if and only if $\lambda\gamma^{d(\overline{H})} \leq \mu$.*

Proof. Sufficiency: If $\lambda\gamma^{d(\overline{H})} \leq \mu$, then $\mathbf{x}^* = \mathbf{x}^0$.

Recall the definition of the densest subgraph V.5. With $\gamma > 1$, $\lambda\gamma^{d(H(\mathbf{x}))} \leq \lambda\gamma^{d(\overline{H}(\mathbf{x}))} \leq \mu$ for all possible induced subgraphs in G . This means that there is no subgraph, $H \in \mathcal{H}$, for which $\lambda\gamma^{d(H)} > \mu$. We can conclude that $\mathbf{x}^* = \mathbf{x}^0$ using the contrapositive of Theorem V.6: If there is no subgraph $H \in \mathcal{H}$ with density $d(H)$ for which $\lambda\gamma^{d(H)} > \mu$, then $\mathbf{x}^* = \mathbf{x}^0$.

Necessity: If $\mathbf{x}^* = \mathbf{x}^0$, then $\lambda\gamma^{d(\overline{H})} \leq \mu$.

The result follows from the contrapositive of Theorem V.6: If $\mathbf{x}^* = \mathbf{x}^0$, then there is no subgraph $H \in \mathcal{H}$ with density $d(H)$ for which $\lambda\gamma^{d(H)} > \mu$. Therefore, all induced subgraphs, including the densest subgraph have density for which $\lambda\gamma^{d(H)} \leq \mu$. □

Appendix E: Proof for Corollary V.10

Corollary. *If G is a k -regular, complete multipartite, or complete multipartite with k -regular islands network, then $\overline{H} = G$. That is, for these structured networks, the densest subgraph*

is the overall graph.

Proof. We proved previously in [5] that the solution of the Most-Probable Configuration Problem for any parameters $\left(\frac{\lambda}{\mu}, \gamma\right)$ in Regime II) **Endogenous Infection Dominant**: $0 < \frac{\lambda}{\mu} \leq 1, \gamma > 1$, over k -regular, complete multipartite, complete multipartite with k -regular islands networks is either \mathbf{x}^0 and/or \mathbf{x}^N ; the solution to the Most-Probable Configuration Problem for these networks is not a non-degenerate configuration in Regime II). We will use this and Corollary V.8 to prove this corollary.

Consider the contrapositive of Corollary V.8: Let the density of the network be $d(G) = \frac{E}{N}$. Then, the most-probable configuration is not a non-degenerate configuration, $\mathbf{x}^* \in \{\mathbf{x}^0, \mathbf{x}^N\}$, if and only if there does not exist any subgraph $H \in \mathcal{H}$ with density $d(H) = \frac{E'}{N'}$ for which $\lambda\gamma^{d(H)} > \mu$, or

$$\frac{\log\left(\frac{\lambda}{\mu}\gamma^{d(G)}\right)}{\log\left(\frac{\lambda}{\mu}\gamma^{d(H)}\right)} < \frac{N'}{N}.$$

This implies that all the induced subgraphs, $H \in \mathcal{H}$, in networks whose solution to the Most-Probable Configuration Problem is not a non-degenerate configuration in Regime II), satisfy the condition that $\lambda\gamma^{d(H)} \leq \mu$ or

$$\frac{\log\left(\frac{\lambda}{\mu}\gamma^{d(G)}\right)}{\log\left(\frac{\lambda}{\mu}\gamma^{d(H)}\right)} \geq \frac{N'}{N},$$

for all $0 < \frac{\lambda}{\mu} \leq 1, \gamma > 1$.

Depending on the effective infection rate and the endogenous infection rate, $\left(\frac{\lambda}{\mu}, \gamma\right)$, the first condition $\lambda\gamma^{d(H)} \leq \mu$ may not be satisfied. However, since $\frac{N'}{N} \leq 1$ regardless of the parameters and the underlying network, the second condition is satisfied if

$$\frac{\log\left(\frac{\lambda}{\mu}\gamma^{d(G)}\right)}{\log\left(\frac{\lambda}{\mu}\gamma^{d(H)}\right)} \geq 1, \quad \forall H \in \mathcal{H}.$$

Since $\gamma > 1$, this means that $d(H) \leq d(G)$ for all possible induced subgraph. As this only depend on the structure of the underlying network, we can conclude that $d(H) \leq d(G)$ for networks whose most-probable configuration can only be \mathbf{x}^0 and/or \mathbf{x}^N .

□

[1] M. Newman, *Networks: an Introduction* (Oxford University Press, 2010).

- [2] P. Csermely, A. London, L.-Y. Wu, and B. Uzzi, *Journal of Complex Networks* **1**, 93 (2013).
- [3] R. Pastor-Satorras and A. Vespignani, *Physical Review E* **65**, 035108 (2002).
- [4] D. R. De Souza and T. Tomé, *Physica A: Statistical Mechanics and its Applications* **389**, 1142 (2010).
- [5] J. Zhang and J. M. F. Moura, *IEEE Journal of Selected Topics in Signal Processing* **8**, 537 (2014).
- [6] J. Zhang and J. M. F. Moura, in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2013) pp. 5411–5414.
- [7] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos, in *Proceedings of the International Symposium on Reliable Distributed Systems* (Florence, Italy, 2003) pp. 25–34.
- [8] M. R. Weeks, S. Clair, S. P. Borgatti, K. Radda, and J. J. Schensul, *AIDS and Behavior* **6**, 193 (2002).
- [9] D. J. Watts and S. H. Strogatz, *Nature* **393**, 440 (1998).
- [10] G. R. C. Godsil, *Algebraic Graph Theory* (Springer-Verlag, 2001).
- [11] M. Draief, A. Ganesh, and L. Massoulié, in *Proceedings of the International Conference on Performance Evaluation Methodologies and Tools* (ACM, 2006) p. 51.
- [12] A. Ganesh, L. Massoulié, and D. Towsley, in *Proceedings of the Annual Joint Conference of the IEEE Computer and Communications Societies* (Miami, USA, 2005) pp. 1455–1466 vol. 2.
- [13] J. R. Norris, *Markov Chains*, 2008 (Cambridge University Press, 1998).
- [14] F. P. Kelly, *Reversibility and Stochastic Networks* (Cambridge University Press, 2011).
- [15] S. Bonaccorsi, S. Ottaviano, F. De Pellegrini, A. Socievole, and P. Van Mieghem, *Physics Review E* **90**, 012810 (2014).
- [16] A. Billionnet and M. Minoux, *Discrete Applied Mathematics* **12**, 1 (1985).
- [17] E. Boros and P. L. Hammer, *Discrete Applied Mathematics* **123**, 155 (2002).
- [18] M. Grötschel, L. Lovász, and A. Schrijver, *Combinatorica* **1**, 169 (1981).
- [19] L. Lovász, in *Mathematical Programming The State of the Art* (Springer, 1983) pp. 235–257.
- [20] A. Krause, *The Journal of Machine Learning Research* **11**, 1141 (2010).
- [21] J. Leskovec and A. Krevl, “SNAP Datasets: Stanford large network dataset collection,” <http://snap.stanford.edu/data> (2014).
- [22] S. P. Borgatti, in *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers* (National Academies Press, 2003) p. 241.

- [23] J. Zhang and J. M. F. Moura, in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2014) pp. 1125–1129.
- [24] S. Khuller and B. Saha, in *Automata, Languages and Programming* (Springer, 2009) pp. 597–608.
- [25] S. Wasserman, *Social Network Analysis: Methods and Applications*, Vol. 8 (Cambridge university press, 1994).
- [26] S. P. Borgatti and M. G. Everett, *Social Networks* **21**, 375 (2000).
- [27] U. Brandes, J. Pfeffer, and I. Mergel, *Studying Social Networks: A Guide to Empirical Research* (Campus, 2013).